AFRL-HE-AZ-TR-2006-0046

Automated Communications Analysis System using Latent Semantic Analysis

Peter W. Foltz

New Mexico State University
Department of Psychology
P.O. Box 30001, MS 3452
Las Cruces NM 88003

August 2006
Final Report for Nov 2003 to January 2005

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Readiness Research Division

NOTICES

This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its idea or findings.

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner, licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

This report has been reviewed by the Public Affairs office and is considered cleared for public release.

Direct requests for copies of this report to: http://www.dtic.mil

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-AZ-TR-2006-0046

This technical report has been reviewed and is approved for publication.

//Signed

WINSTON BENNETT, JR.
Contract Monitor

//Signed

HERBERT H. BELL
Technical Advisor

//Signed

DANIEL R. WALKER, Colonel, USAF
Chief, Warfighter Readiness Research Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
August 2006	Final	17 Nov 03 to 17 Jan 05

4. TITLE AND SUBTITLE

Automated Communications Analysis System using Latent Semantic Analysis

5a. CONTRACT NUMBER
FA8650-03-1-6382

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)

Peter W. Foltz

5d. PROJECT NUMBER
1123

5e. TASK NUMBER
AS

5f. WORK UNIT NUMBER
01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

New Mexico State University
Department of Psychology
P.O. Box 30001, MS 3452
Las Cruces NM 88003

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Training Research Division
6030 South Kent Street
Mesa AZ 85212-6061

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL; AFRL/HEA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-AZ-TR-2006-0046

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
This research was conducted under the Historically Black Colleges and Universities/Minority Institutions (HBCU/MI) Program.

14. ABSTRACT

This contract effort was initiated to develop and evaluate automated methods for analyzing verbal protocols related to air combat missions and verbal interactions amongst pilots during a mission and during the debriefing process to assess knowledge proficiency. In this report, the contractor describes prior research on communication analysis and how it can inform assessment of individual and team cognitive processing. Then, they describe techniques using Latent Semantic Analysis (LSA) which can perform analyses of communications and provide automated assessment of this rich source of data. Finally, they propose a course of research to evaluate LSA's effectiveness as a software agent to monitor communications.

15. SUBJECT TERMS
Automated communications; behavioral measures; cognitive measures; embedded assessment technology; HBCU/MI; human learning; synthetic task environments; verbal interactions

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Dr Winston Bennett
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR	24	**19b. TELEPHONE NUMBER** *(include area code)*

PREFACE

This research was performed for the Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Readiness Research Division (AFRL/HEA) under Air Force Contract FA8650-03-1-6382, Workunit No. 1123-AS-01, Automated Communications Analysis System using Latent Semantic Analysis. The Laboratory Principal Investigator was Dr Winston Bennett. The effort was performed under the Historically Black Colleges and Universities/Minority Institutions (HBCU/MI) Program.

The contractor research team consisted of Dr. Peter W. Foltz, Department of Psychology, New Mexico State University, and graduate students, Melanie Martin, Computer Science, and Ahmed Abdelali, Computer Research Laboratory.

This research was completed in collaboration with the Cognitive Engineering Research on Team Tasks (CERTT) laboratory including, Nancy Cooke, Preston Kiekel, Jamie Gorman, Susan Smith, and David Farwell from the Computing Research Laboratory.

CONTENTS

AUTOMATED COMMUNICATIONS ANALYSIS SYSTEM USING LATENT SEMANTIC ANALYSIS

Background

Assurance that personnel have acquired the appropriate knowledge and competencies is a critical aspect of training. Within Distributed Mission Training (DMT) and in live -flight exercises, assessment must often rely on an analysis of verbal communication generated either during missions or from the debriefing. This communication data provides a rich indication of cognitive processing at both the individual and the team level and can be tied back to an individual team member's abilities and knowledge. However, due to the volume of data and the paucity of automated methods, such analyses have thus far been difficult to perform in real time.

In the present research, automated techniques to analyze verbal communications from simulated flight exercises were be developed and evaluated. These techniques were primarily based on Latent Semantic Analysis (LSA), an artificial intelligence technology that permits characterization of the semantic content in language. The analysis of verbal communication techniques were to be evaluated on their ability to identify knowledge proficiencies based on cognitive and behavioral measures and to provide remediation of knowledge gaps. These techniques can be applied to a wide range of settings in the Air Force for the monitoring and analysis of communications. A proof-of-concept demonstration was developed and a feasibility study was conducted to evaluate the development of an operational, real -time, communication assessment system.

In this report we describe prior research on communication analysis and how it can inform assessment of individual and team cognitive processing. Then, we describe techniques using LSA which can perform analyses of communications and provide automated assessment of this rich source of data. Finally, we propose a course of research to evaluate LSA's effectiveness as a software agent to monitor communications.

Monitoring Verbal Interactions

Verbal communications provide a rich source of data, incorporating both information about the content of the communications and the patterns of communications. While the analysis of the content provides a rich characterization of the knowledge, skills, and verbal abilities of people, it has been a time-consuming and difficult task to perform. Therefore, in the past, observations of individual and team communication have largely been quantified in terms of overall communication frequency or frequencies of specific communications acts (e.g., acknowledgment, question, planning). Results using such frequencies have been mixed. Transcription and coding processes are very tedious and costly. In addition, information regarding the sequential patterns of communication and the flow of communication among team members has received little attention (see Bowers, Jentsch, Salas, & Braun, 1998). In general, research on assessing individual and team competencies in real-time is hindered by the paucity of methods and tools for measuring verbal communication in a cost-effective way (i.e., automated analyses, task-embedded). Before addressing our approach to this problem, we provide some background on theories, methods, and empirical findings relevant to analyses of communication.

Most commonly, analyses of communication data have either focused on low-level quantitative measures such as duration of communication or on encoding the communication into prescribed content categories (Contractor & Grant, 1996). The former approach can be used to capture some of the complexity of communication patterns through time (usually operationalized with physical measures) by modeling the quantitative measures using lag sequential and/or Markov chains, time series modeling, Fourier analysis (Watt & VanLear, 1996, p. 12), or other methods that uncover the unfolding of patterns over time (Sanderson & Fisher, 1994). The latter approach involves first selecting a coding scheme that includes all interesting categories of communication meaning, such as the rules being displayed in the conversation, the types of speech, or the actual meaning of the discussion. The transcribed discourse is then divided into the smallest units of meaning, then those pieces of text that correspond to the categories of interest are tallied (Emmert & Barker, 1989). Communication patterns can be analyzed either as frequency counts of the categories or as a series of events (called "interaction analysis", see Emmert, 1989, for discussion; Poole, Holmes, Watson, & DeSanctis, 1993, for an example), using lag sequential analysis or other tools (see Holmes, 1997, for an example).

Quantitative and content -based approaches have their own merit--and their own costs. For the content-based approach, multiple coders are intensively trained, and must have adequate agreement. Emmert and Barker (1989) cite an example of a study requiring 28 hours of transcription and encoding for each hour of communication (p. 244). But the advantage is that communication content is captured, including in some cases, nonverbal communication (Donaghy, 1989). More quantitative approaches are somewhat easier in data collection (although speaker, listener, and communication duration is often tedious to transcribe from audio tape), but fail to capture meaning (see Contractor & Grant, 1996, for an exception, in which agreement between communicators is modeled with a numeric value). Both approaches have been used to analyze communication among groups of larger than two, but the transcription and encoding tasks become even more cumbersome as the complexity of the communication and the possibility for parallel audio streams increases.

In summary, there is a general consensus that continuous streams of rich data are necessary to describe the unfolding process of communication, but that automatic methods for doing this are nonexistent or problematic (Smith, 1994). Even automatic collection of event data at a computer is currently ineffective. Automatic collection of group process behavior not related to the computer is currently unavailable. If researchers are interested in modeling who talks to whom and for how long, human raters must record and time-stamp these data. Communication content is even more labor intensive, since it requires that human raters classify the discourse into prescribed categories. There are currently no automatic methods for doing this. Nonetheless, findings from team communication studies in which manual transcription and coding have been used appear promising.

Team Communication

Similar to the methods used to analyze more general communication, team communication has largely been quantified in terms of overall communication frequency (and sometimes rate of communication or frequency wit h which a team member initiates communication; Oser, Prince,

Morgan, & Simpson, 1991), and frequencies of specific communications acts (e.g., acknowledgment, question, planning). In terms of overall frequency, results have been equivocal. In some cases studied, high performing teams communicate with higher overall frequency than low performing teams (Foushee & Manos, 1981; Mosier & Chidester, 1991; Orasanu, 1990), but in other cases this finding has not been supported (e.g., Thornton, 1992). Some studies indicate that overall communication frequency is reduced during high workload periods (Kleinman & Serfaty, 1989; Oser, et al., 1991), whereas others indicate increases in communication frequency under relatively high workload (e.g., Stout, 1995). Some of these differences may be due to other factors such as the task or the nature of the teams. For example, Bowers, Urban, and Morgan (1992) found that the correlation between communication frequency and team performance was tied to whether the team was hierarchical in structure. In other cases, mixed results may be due to the use of frequency measures devoid of communication content or sequential information.

Communication content associated with team studies has been analyzed by transcribing the audio information and segmenting it into units associated with speech turns or complete thoughts. Then the segmented transcript is coded using categories pertinent to the hypothesis or research problem. Some examples of content categories include, speech acts such as acknowledgments, requests, statements, or answers to questions; errors such as violation in standard format, and use of terminology such as standard military terms. Results tend to be more specific and of greater practical significance than those associated with frequency analyses. For instance, Achille, Schulze, and Schmidt-Nielsen (1995) found that the use of military terms, acknowledgments, and identification statements increased with experience. Similarly Jentsch, Sellin-Wolters, Bowers, and Salas (1995) found that faster teams made more leadership statements and more observations about the environment than slower teams. In addition, the communication of faster teams was more standard five minutes before the problem than for slower teams.

Parallel to general trends in communication analysis, recent research also suggests that advances in team communication analysis and understanding may come from extending analysis beyond single dimensions such as frequency of content category to more complex patterns, taking into account multiple dimensions including content, frequency, sequence, and communication flow. For instance, Bowers and colleagues (1998) analyzed the sequence of content categories occurring in communication in a flight simulator task. They found that high team effectiveness was associated with consistent responding to uncertainty, planning, and fact statements with acknowledgments and responses in comparison to lower performing teams. Similarly, Bowers, Braun, and Kline (1994) found that a two-category sequence was superior to simple frequencies at predicting performance on an aerial reconnaissance task. On the basis of results like these, Salas, Bowers, and Cannon-Bowers (1995) conclude that "It is likely that additional pattern-based analyses will emerge in future literature as a means to understand the impact of communication on team performance" (p. 64).

In summary, recent research on team communication that takes analysis beyond overall frequencies, to explore content and sequential information, is sparse, but shows much promise. A major stumbling block in this kind of research is the costliness of manual analysis needed to code content and transcribe sequential and pattern information from audio records (Achille et al.,

1995). Salas et al. (1995), highlight this research need and state that "… methods to interpret team process information, which until now has been almost exclusively a manual task, would benefit from automation" (p. 69). Indeed, team cognition work in general is hampered by the paucity of automated methods and data collection limits. The objective of the effort described in this proposal is to develop and evaluate automated methods that address this problem.

Methodological Background

Latent Semantic Analysis. One solution to the analysis of team communications is the use of LSA. LSA is a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse. It is not a traditional natural language processing or artificial intelligence program; it uses no humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like, and takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs.

The primary assumption of LSA is that there is some underlying or "latent" structure in the pattern of word usage across contexts (e.g., paragraphs or sentences within texts), and that statistical techniques can be used to estimate this latent structure. Through an analysis of the associations among words and contexts, the method produces a high-dimensional representation in which words that are used in similar contexts will be represented as being more semantically associated. Using this representation, words, sentences, or larger units of text may be compared against each other to determine their semantic relatedness. A brief overview of the technical approach to applying LSA is described here. Additional details may be found in Berry (1992, Deerwester, Dumais, Furnas, Landauer & Harshman (1990), Landauer & Dumais (1997), Landauer, Foltz & Laham (1998).

To analyze a text or texts, LSA first generates a matrix of occurrences of each word in each context (e.g., sentences or paragraphs). In this pre-processing stage, each cell of the matrix contains a transformation of the frequency of the occurrences of each word. This transformation typically used is the log of the frequency of the word times the entropy of its frequency across all contexts. Transforms of this or similar kinds have long been known to provide marked improvement in information retrieval (Harman, 1986) and have been found important in several applications of LSA. The transforms are important for correctly representing a passage as a combination of the words it contains because they emphasize specific meaning-bearing words.

LSA then applies singular-value decomposition (SVD), a form of factor analysis, or more properly the mathematical generalization of which factor analysis is a special case. The SVD scaling decomposes the word-by-context matrix into a set of, typically 100 to 300, orthogonal factors (or dimensions) from which the original matrix can be approximated by linear combination. Instead of representing contexts and terms directly as vectors of independent words, LSA represents them as continuous values on each of the orthogonal indexing dimensions derived from the SVD analysis. Since the number of factors or dimensions is much smaller than the number of unique terms, words will not be independent. For example, if two terms are used in similar contexts, they will have similar vectors in the reduced-dimensional LSA

representation. One advantage of this approach is that matching can be done between two pieces of textual information, even if they have no words in common.

One can interpret the analysis performed by SVD geometrically. The result of the SVD is a -dimensional vector space containing a vector for each term and each document. The location of term vectors reflects the correlations in their usage across documents. Similarly, the location of document vectors reflects correlations in the terms used in the documents. In this space the cosine or dot product between vectors corresponds to their estimated semantic similarity. Thus, by determining the vectors of two pieces of textual information, we can determine the semantic similarity between them.

The number of dimensions retained in LSA is an empirical issue. but rather, An optimal dimensionality should be found that will cause correct induction of underlying relations because the underlying principle is that the original data should not be perfectly regenerated. The customary factor-analytic approach of choosing a dimensionality that most parsimoniously represents the true variance of the original data is not appropriate. Instead some external criterion of validity is sought, such as the performance on a synonym test or prediction of the missing words in passages if some portion is deleted in forming the initial matrix.

LSA's performance has been evaluated as a representational model and measure of human verbal concepts as well as has been used for a wide variety of applications that require the analysis of the conceptual content of textual material. LSA's performance has been assessed more or less rigorously in several ways: (a) as a predictor of query-document topic similarity judgments in information retrieval (Deerwester et al., 1990); (b) as a simulation of agreed upon word-word relations and of human vocabulary test synonym judgments (Landauer & Dumais, 1997), (c) as a simulation of human choices on subject-matter multiple choice tests, (d) as a predictor of text coherence and resulting comprehension (Foltz, Kintsch & Landauer, 1998), (e) as a simulation of word-word and passage-word relations found in lexical priming experiments (Landauer & Dumais, 1997), (f) as a predictor of subjective ratings of text properties, i.e., grades assigned to essays (Rehder et al.,1998; Foltz, 1996; Foltz, Laham & Landauer, 1999), and (g) as a predictor of appropriate matches of instructional text to learners essays (Wolfe et al., 1998).

While assessing the performance of LSA, the above tests have also permitted the derivation of applications that incorporate LSA for measuring the conceptual content of textual information. Existing applications have included, information retrieval ad filtering programs, techniques for automatically scoring and commenting essays, methods determining the appropriate training material for individual learners, and methods for matching. In this project, we employed similar approaches to analyze and categorize the discourse of team communication.

Program Plan

Objectives

The main objective was to develop and evaluate techniques for the analysis of communication data that could be incorporated into a LSA-based software agent to monitor free-form verbal interactions. Because the techniques will be automated, they can be more cost-effective than the

traditional manual methods. The techniques should ultimately facilitate the development of systems for automated real-time assessment and diagnosis of knowledge and competencies. We capitalized on the specific capabilities of our research skills to perform analysis using Latent Semantic Analysis, to assess knowledge, and to work with communications data.

There were four primary tasks associated with the project. The tasks are touched upon briefly in this report along with the primary objectives associated with the tasks. Additional details on the tasks and objectives are provided below in the technical discussion. It should be noted that because of the short scope of this research grant and restricted funding, the data obtained were collected and part of the analyses were performed concurrently with ongoing communication analysis research funded through the Office of Naval Research (ONR).

TASK 1: Collect and assess verbal interaction data from DMT during live-flight activities on instrumented ranges

Objectives: Obtain transcripts and associated performance data from Ai r Force DMT exercises. Conduct field research and work closely with Air Force personnel to obtain appropriate data and determine the efficacy of evaluating existing data. Because of the short duration of this research project and since DMT data were not readily available, data collected from simulated Uninhabited Aerial Vehicles (UAV) missions conducted in the Cognitive Engineering Research on Team Tasks (CERTT) laboratory from New Mexico State University (NMSU) and Arizona State University (ASU) were used to test methods.

TASK 2: Develop, iteratively refine, and evaluate methods for the analysis of DMT verbal data and/or other available verbal data

Objectives: Test and develop natural language monitoring techniques and real-time language-driven data assimilation and analysis tools to support training and rehearsal. Use methods that have already been developed at NMSU as well as develop new methods, test them on CERTT UAV. Evaluate performance of methods both individually and in combination. Develop a proof-of-concept system that demonstrates automated assessment of transcripts.

TASK 3: Associate discourse content with cognitive and behavioral measures that together are diagnostic knowledge proficiencies.

Objectives: Use additional performance data and/or knowledge proficiency measures obtained from tasks, and develop methods to predict the cognitive and behavioral measures based on the discourse content. Based on a task analysis of the task, tie the predicted knowledge proficiency to automated text-based feedback.

TASK 4: Prepare Final Report

Objectives: Develop a report on the findings and methodological details on the assessment methods developed. Investigate and report on the feasibility of developing an operational software agent that automatically analyzes field data and communications.

Technical Details

Three corpora of team transcripts were collected as a result of three different team experiments that simulate flight of a UAV in the CERTT Lab's synthetic task environment (CERTT UAV-STE). CERTT's UAV-STE is a three-team member task in which each team member is provided with distinct, though overlapping, training; has unique, yet interdependent roles; and is presented with different and overlapping information during the mission. The overall goal is to fly the UAV to designated target-areas and to take acceptable photos at these areas. To complete the mission, the three team members need to share information with one another and work in a coordinated fashion. Most communication is done via microphones and headsets, although some involves computer messaging.

The three corpora are labeled by experiment name: AF1, AF3, and AF4. Each corpus consists of a number of team-at-mission transcripts, where mission duration is approximately 40 minutes. Some statistics are shown in Table 1. All communication was manually transcribed. Some team - at-missions had to be excluded due to recording and transcription difficulties.

Table 1. Corpora Statistics

Corpus	Transcripts	Teams	Missions	Utterances
AF1	67	11	7	20245
AF3	85	21	7	22418
AF4	85	20	5	22107

Description of Semantic Spaces

To train LSA we added 2257 documents to the transcripts of each corpus. These documents consisted of training documents and pre- and post-training interviews related to UAVs. We created four semantic spaces: AF1, AF3, AF4, and AF1-3-4 (combines all three corpora and training materials). In each case we used an approximately 300 dimensional semantic space. Unless otherwise noted all results reported were computed using the AF1-3-4 semantic space.

Predicting Team Performance

Throughout the CERTT UAV-STE experiments a performance measure was calculated to determine each team’s effectiveness at completing the mission. The performance score was a composite of objective measures including: amount of fuel/film used, number/type of photographic errors, time spent in warning and alarm states, and unvisited waypoints. This composite score ranged from -200 to 1000. It should be noted that the method for calculating the performance scores was changed between AF1 and the two later experiments. Therefore, results using the performance measures cannot be compared between AF1 and the other two experiments. The score is highly predictive of how well a team succeeded in accomplishing their mission. We used two approaches to predict these overall team performance scores: correlating entire mission transcripts with one another and by correlating tag frequencies with the scores.

Prediction Using Whole Transcripts

Our first approach to measuring content in team discourse is to analyze the transcript as a whole. Using a k-nearest neighbor method that has been highly successful for scoring essays with LSA (Landauer et al., 1998), we used whole transcripts to predict the team performance score. The predicted team performance s cores was as follows: Given a subset of transcripts, S, with known performance scores, and a transcript, t, with unknown performance score, we can estimate the performance score for t by computing its similarity to each transcript in S. The similarity between any two transcripts is measured by the cosine between the transcript vectors in the semantic space. To compute the estimated score for t, we take the average of the performance scores of the 10 closest transcripts in S, weighted by cosines. A holdout procedure was used in which the score for a team's transcript was predicted based on the transcripts and scores of all other teams (i.e., a team's score was only predicted by the similarity to other teams). Tests on the AF1 corpus showed that the LSA estimated performance scores correlated strongly with the actual team performance scores ($r = 0.76$, $p < 0.01$, $r = 0.63$, $p < .01$) when correcting for the repeated measure structure (see Figure 1 and Martin & Foltz, 2004). Thus, the results indicate that we can accurately predict the overall performance of the team (i.e., how well they fly and complete their mission) just based on an analysis of their transcript from the mission.

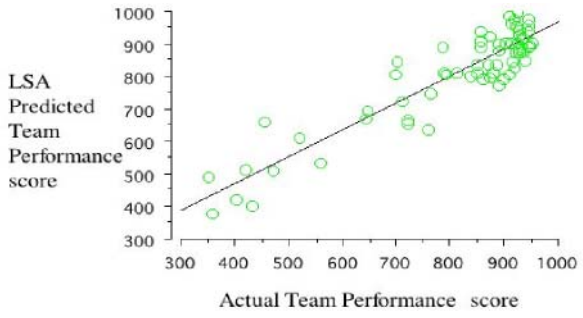

Figure 1. Correlation: Predicted and Actual Team Performance for AF1.

Generalization of Team performance scores for different corpora

While the results were successful for the AF1 corpus, it is important to determine if similar results can be found for the other two corpora. In addition, it is important to determine if the algorithm can operate successfully by training the algorithm on the performance scores of one corpus in order to predict performance scores on another corpus. This approach would be equivalent to having collected N transcripts from teams flying UAVs on a set of particular missions and then trying to predict a new set of teams performing a different set of missions. Thus, the generalization test, will determine how robust such a system could be in more realistic contexts where different teams may have to fly entirely novel missions.

We tested the generalization for the AF3 set of transcripts, by training our algorithm on the performance scores of AF3 performance scores and predicting the performance scores from the other experiment (AF4). Using the 10 closest transcripts, as before, the LSA estimated scores strongly correlated with the actual scores or AF3, showing only a 4% degradation in performance (see Table 2.). Thus, there was a high level of generalization from one training corpus to predicting the performance scores of another.

	Training Set		
	AF3	AF4	Difference
AF3	0.72	0.66	-4%

Table 2. Predicted -Actual Score Correlations When Varying the Training Set

Generalization of Semantic Spaces for Whole Transcript Prediction

The above generalization results also raise the issue about whether the size and type of semantic space used is important. For instance, how well do these predictions hold if the semantic space is based on a corpus that does not contain the missions being tested?

To demonstrate the generalization of our algorithm over varying semantic spaces, we compared the correlation of estimated and predicted team scores for AF1 and AF3 transcripts using the AF1-3-4 semantic space. The results, shown in Table 3, confirm that performance is not significantly changed by using a larger, more general, semantic space or even by using other semantic spaces of approximately equivalent size, but not containing the tested missions as part of the semantic space. It further shows that LSA is robust over a range of different sized corpora.

	Semantic Space used			
	AF1	AF3	AF1_3_4	Difference
AF1	0.76		0.77	+1%
AF3		0.75	0.72	-4%

Table 3. Predicted -Actual Scores Correlations When Varying Semantic Spaces

Predicting Workload Level Using Whole Transcripts

In the AF3 and AF4 experiments, workload in the missions was manipulated such that all teams received some missions with double the workload. In AF3, three out of seven missions for each team were high workload and in AF4 one of the five missions for each team was high workload. Using a similar k-nearest neighbor algorithm on whole transcripts to predict workload we found strong correlations between the actual and predicted workloads.

The algorithm first assigns a score of 1 for high workload and 0 for low workload missions. Then it takes the average, weighted by distance in the semantic space, of the ten closest team-at

missions, excluding all missions from the current team and from other experiments. Team-at missions whose weighted average is greater than the cutoff of 0.25 are labeled “high workload,” others are labeled “low workload.” We computed the kappa statistic to assess the agreement between the actual and predicted workload.

AF3: kappa = 0.91
AF4: kappa = 0.84

This result shows that the approach can accurately classify a mission as to whether the team was under high or low workload. Further work is underway to determine the components of discourse that permit the characterization of low or high workload in teams. This work addresses task 3 of the project, showing that we can accurately measure important behavioral components of performance automatically based on communication data.

Automated Discourse Tagging

Our goal is to use semantic content of team dialogues to better understand and predict team performance. The approach we focus on here is to study the dialogue on the turn level. We designed an algorithm to learn from human tagged content of the communication data and then apply the tool to new communication data.

We used the tag-set developed by Bowers et al. (1998) to analyze airplane cockpit team communication. The set consists of tags for acknowledgement, action, fact, planning, response, uncertainty, and non-task communication. The frequency of the occurrence of these tags in team discourse has been shown to be predictive of team performance.

Working within the limitations of the manual annotations, we developed an algorithm to tag transcripts automatically, resulting in some decrease in performance, but a significant savings in time and resources.

We established a lower bounds tagging performance of 0.27 by computing the tag frequency in the 12 AF1 transcripts tagged by two taggers. If all utterances were tagged with the most frequent tag, the percentage of turns tagged correctly would be 27%.

<u>Algorithm for Automatic Annotation</u>

In order to test our algorithm to automatically annotate the data, we computed a "corrected tag" for all 2916 turns in the 12 team-at-mission transcripts tagged by two taggers. This was necessary due to the only mode rate agreement between the taggers. We used the union of the sets of tags assigned by the taggers as the "corrected tag."

The union, rather than the intersection, was used since taggers sometimes missed relevant tags within a turn. The union of tags assigned by multiple taggers better captures all likely tag types within the turn. A disadvantage to using “corrected tags” is the loss of sequential tag information within individual turns. However, the focus of this research was on identifying the existence of relevant discourse, not on its order within the turn.

Then, for each of the 12 team-at-mission transcripts, we automatically assigned "most probable" tags to each turn, based on the corrected tags of the "most similar" turns in the other 11 team-at-missions, using Martin and Foltz (2004) algorithm for LSA and LSA+.

The LSA+ algorithm adds two discourse features to the LSA algorithm: for any turn with a question mark, "?," we increased to probability that uncertainty, "U," would be one of the tags in its predicted tag; and for any turn following a turn with a question mark, "?," we increased to probability that response, "R," would be one of the tags in its predicted tag. Using LSA+ the performance is now only 10% and 15% below human-human agreement, depending on which agreement measure is used (see Table 4).

Annotators-Agreement	C-Value	Kappa
Human-Human	0.70	0.48
LSA-Human	0.59	0.48
LSA+Human	0.63	0.53

Table 4. Kappa and C -Values.

We realize that training our system on tags where humans had only moderate agreement is not ideal. Our failure analyses indicated that the distinctions our algorithm has difficulty making are the same distinctions that the humans found difficult to make, so we believe that improved agreement among human annotators would result in similar improvements for our algorithm. The results suggest that we can automatically annotate team transcripts with tags While the approach is not quite as accurate as human taggers, LSA is able to tag an hour of transcripts in under a minute. As a comparison, it can take half an hour or longer for a trained tagger to do the same task.

Measuring Agreement

The C-value (Schvaneveldt, 1990) measures the proportion of inter-coder agreement, but does not take into account agreement by chance. To adjust for chance agreement we computed Cohen's Kappa (Cohen, 1960), as shown in Table 4.

Predicting Performance from Tags

To relate performance data to the behavioral measures based on the types of communications, we computed correlation s between the team performance score and tag frequencies in each team-at-mission transcript. The tags for all 20545 utterances in the AF 1 transcripts were first gene rated using the LSA+ method. The tag frequencies for each team-at mission transcript were then computed by counting the number of times each individual tag appeared in the transcript and dividing by the total number of individual tags occurring in the transcript.

Our results indicate that frequency of certain types of utterances correlate with team performance. The correlations for tags predicted by computer are shown in T able 5.

We see that the automated tagging provides useful results that can be interpreted in terms of team processes. Teams that tend to state more facts and acknowledge other team members more tend to perform better. Those that express more uncertainty and need to make more responses to each other tend to perform worse. These results are consistent with those found in Bowers et al. (1998), but were generated automatically rather than by the hand-coding done by Bowers.

TAG	PEARSON CORR.	Sig. 2-Tailed
Acknowledgement	0.335	0.006
Fact	0.320	0.008
Response	-0.321	0.008
Uncertainty	-4.460	0.000

Table 5. Tag to Performance Correlations.

Generalization of Tag Prediction

To test the ability of our automatic tagging algorithm to generalize, we trained a new annotator. He was trained on the AF1 corpus and in testing achieved good agreement with the previous annotators: Kappa was 0.72. Given this level of agreement we had him tag 20 randomly selected transcripts from each of AF3 and AF4 (approximately 24% of the total discourse in these corpora). We were then able to compare our automatically predicted tags for AF3 and AF4 to his tags (see Table 6). In this approach, we train the system on the AF 1 tags to determine how well the system can predict the human generated tags on the AF1, AF3 and AF4 corpora.

	AF1	AF3	AF4
Kappa	0.53	0.56	0.54
C-value	0.63	0.66	0.64

Table 6. LSA+ - Annotator Agreement

The results indicate that humans can consistently, although not highly accurately, use the Bowers tag set across the three corpora, and that the LSA+ algorithm can consistently predict the tags. As with the whole transcript prediction we were able to show generalization across semantic spaces: training on the tags in AF1 to predict tags in AF1, produced equivalent kappas (to two decimal places) using the AF1 and AF1 -3-4 semantic spaces.

In addition we varied the set of tags used for training. In the AF1-3-4 semantics space, predicting tags for the AF3 corpus showed only a 5% degradation in performance when the system was trained on the AF1 tags rather than on the AF3 tags. We believe this demonstrates the robustness and ability to generalize, at least within the UAV-STE domain, of the LSA+ algorithm.

Development of a demonstration prototype

As part of this project, a web-based demonstration system was developed that could take incoming transcripts of teams and generated automated performance scores. A screen shot of the system is shown in Figure 2. It illustrates the output of the analysis of a transcript displaying a

number of LSA and other statistics that can be useful for characterizing the quality of the team's performance. In addition to basis statistics about the transcript as a whole, it computes the frequencies of the predicted tags. In the discourse section, the predicted tags, their certainty, coherence with the next turn, and vector length (measure of information content of the turn) are shown next to the discourse.

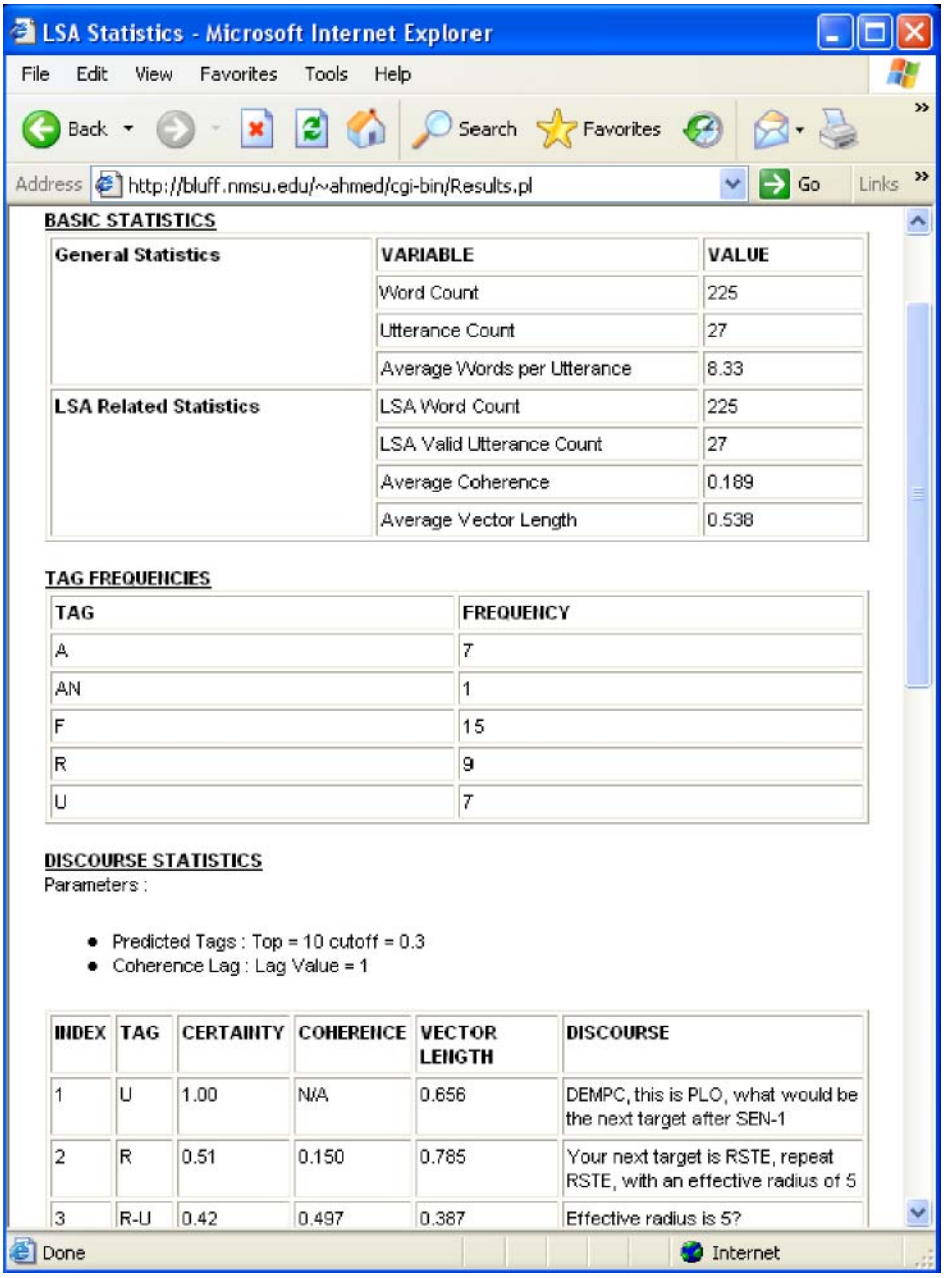

Figure 2. System Screen Shot

Conclusions and Implications

Real time assessment of knowledge and competencies based on communication data has been limited by the methods available to collect and assess this rich data source. Prior attempts at coding for content have relied on tedious hand-coded techniques or have used limited data such as frequencies and durations of communications. With the advent of artificial intelligence techniques that can measure the semantic content of communication discourse, novel methods for the analysis of communication can be applied.

Overall, the results of the study show that LSA-based algorithms can be used for tagging content as well as predicting team performance based on team dialogues. These results extend prior studies and show that the approach is generalizable and not due to specific corpora, semantic spaces, or training sets. Results from the tagging portion of the research are comparable to other efforts of automatic discourse tagging using different methods and different corpora (Stolcke et al., 2000), which found performance within 15% of the performance of human taggers. Unlike the previous efforts though, LSA relies only on a semantic model, ignoring word order and other syntactic and discourse factors. It should be noted that we don't think that LSA is a complete solution to discourse prediction or annotation. It is anticipated that incorporating additional methods that account for syntax and discourse turns should further improve the overall performance, see also Serafin et al. (2004).

In addition to being able to use the LSA-based approach to discourse tagging, this research demonstrates how it can be applied as a method for doing automated measurement of team performance. The LSA-predicted team performance scores correlated strongly with the actual team performance measures. This demonstrates that analyses of discourse can automatically measure how well a team is performing on a mission. This has implications both for automatically determining what discourse characterizes good and poor teams as well as developing systems for monitoring team performance in near real-time.

For example, we can now locate utterances in the semantic space that correspond to places where teams received high or low team scores. These can provide indications of the type of language that is strongly correlated with good and poor performance. It can further identify potential knowledge gaps in teams. Because of the highly interactive nature of the task, there are certain pieces of knowledge that must flow between team members at critical points. These techniques can identify if this information has been conveyed.

In terms of applying this research to team dialogues, the automated methodologies for the analysis of communications data provide cost-effective and efficient approaches for analyzing communications data within DMT environments. Although this research used typed transcripts, Foltz, Laham, and Derr (2003) showed that LSA predictions derived from Automated Speech Recognition output was highly robust. With 40% word error rates, LSA's prediction ability decreased by only 10-15%. Thus, these methods can yield information on team communication patterns that are valid, reliable, and useful to the assessment and understanding of team performance and cognition--necessary prerequisites to the development of team training programs and the design of technologies that facilitate team performance. Some potential applications include:

- evaluating teams or individuals on the basis of communication
- assessment of training programs
- identifying critical events (e.g., loss of situation awareness) and diagnosing team performance based on communication patterns
- cognitive and communication training

In particular, application domains that are communications-intensive and that require a high degree of team coordination can especially benefit from these streamlined methods for assessing team communication.

Research into team discourse is a new but growing area. However, up to recently, the large amounts of transcript data have limited researchers from performing analyses of team discourse. The results of this research show that applying artificial intelligence (AI) and neuro-linguistic programming (NLP) techniques to team discourse can provide accurate predictions of performance. These automated tools can help inform theories of the nature of communication in team performance and also aid in the development of more effective automated team training systems.

LSA provides a basis for the development of tools to measure free-form verbal interactions among team members. Because it can measure and compare the semantic information in these verbal interactions, it can be used to characterize the quality of information expressed. This can be used to determine the knowledge and competencies of personnel engaged in distributed mission training. By linking the results of the LSA analysis to behavioral and cognitive measures, methods can be developed to provide measures of the quality of a person's expertise as well as to identify important gaps in his or her knowledge. Because LSA is automatic, once the data are transcribed, analyses can be performed within seconds or minutes, rather than the weeks or months seen in hand-coded methods.

Assessment for combat mission readiness is a critical training issue. Techniques developed in this proposal can be applied across a wide range of training domains. Along with assessing readiness, the techniques can be used as independent validating measures for evaluating training effectiveness. From an applications-oriented perspective, this research will lead to cost-effective and efficient methods for collecting and analyzing communications data. Additionally, these methodologies may be used to facilitate communication analysis in a host of applied settings including the assessment of teams in air combat command, within Advanced Distributed Learning (ADL) based training, and in command, control, communications, and intelligence (C^3I) centers.

Cited and Relevant References

Achille, L. B., Schulze, K. G., & Schmidt-Nielsen, A. (1995). An analysis of communication and the use of military terms in Navy team training. *Military Psychology, 7*, 95-107.

Beebe, S. A., & Masterson, J. T. (1997). *Communicating in Small Groups.* (5th Ed). New York, NY: Longman.

Berry, M. W. (1992). Large scale singular value computations. *International Journal of Supercomputer Application*s, 6, 13-49.

Bowers, C. A., Braun, C. C., & Kline, P. B. (1994). Communication and team situational awareness. In R.D. Gilson, D. J. Garland, and J. M. Koonce (Eds.), *Situational awareness in complex systems* (pp. 305-311). Daytona Beach, FL: Embry Riddle Aeronautical University Press.

Bowers, C. A., Jentsch, F., Salas, E., & Braun, C. C. (1998). Analyzing communication sequences for team training needs assessment. *Human Factors, 40,* 672-679.

Bowers, C. A., Urban, J. M., & Morgan, B. B., Jr. (1992*). The study of crew coordination and performance in hierarchical team decision making* (Rep. No. TR-92-01). Orlando, FL: University of Central Florida Team Performance Laboratory.

Cannon-Bowers, J. A., & Salas, E. (1998). *Making decisions under stress: Implications for individual and team training.* Washington, D. C.: American Psychological Association.

Cohen, J.A. (1960). *A coefficient of agreement for nominal scales*. Educational Psych Meas, *20*, 37-46

Contractor, N. S., & Grant, S. J. (1996). The emergence of shared interpretations in organizations: A self-organizing systems perspective. In J. H. Watt and A.C. VanLear (Eds.), *Dynamic patterns in communication processes (*pp. 215-230). Thousand Oaks, CA: Sage Publications.

Cooke, N. J. & Gillan, D. J. (1999). Representing user behavior in human-computer interaction. In A. Kent and J. G. Williams (Eds.), *Encyclopedia of Computer Science and Technology (*pp. 283-308). New York: Marcel Dekker, Inc. Also to be reprinted in the *Encyclopedia of Library and Information Science.*

Cooke, N. J., Neville, K. J., & Rowe, A. L. (1996) Procedural network representations of sequential data. *Human-Computer Interaction, 11*, 29-68.

Daft, R. L. & Lengel, R. H. (1986). Organizational information requirements, media richness, and structural design. *Management Science,* 32, 554-571.

Daft, R. L., Lengel, R. H., & Trevino, L. K. (1987). Message equivocality, media selection, and manager performance: Implications for information support systems. *MIS Quarterly, 11,* 355-366.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society For Information Science, 41*, 391-407.

Donaghy, W. C. (1989). Nonverbal communication measurement. In P. Emmert and L. L. Barker (Eds.), *Measurement of Communication Behavior (* pp. 296-332). White Plains, NY: Longman, Inc.

Emmert, P., & Barker, L. L. (1989). *Measurement of Communication Behavior.* White Plains, NY: Longman, Inc.

Emmert, V. J. (1989). Interaction analysis. In P. Emmert & L. L. Barker (Eds.). *Measurement of Communication Behavior (* pp. 218-248). White Plains, NY: Longman, Inc.

Foltz, P. W. (1996) Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28,* 197-202.

Foltz, P. W. (1999). Analyzing survey results with Latent Semantic Analysis. Talk presented at *Society for Computers in Psychology,* Los Angeles, CA.

Foltz, P. W., Kintsch, W. & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes, 25,* 285-307.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to educational technology. *Interactive Multimedia Education Journal of Computer Enhanced Learning. 1* (2).

Foltz, P. W., & Wells, A. D. (1999). Automatically deriving reader's knowledge structures from texts. *Behavior Research Methods, Instruments and Computers, 25,* 208-214.

Foushee, H. C., & Manos, K. (1981). Information transfer within the cockpit: Problems in intracockpit communications. In C. E. Billings and E. S. Cheaney (Eds.). Information transfer problems in the aviation system (Report No. NASA TP-1875). Moffett Field, CA: NASA-Ames Research Center.

Freeman, J.T., Thompson, B.T., & Cohen, M.S. (1999). Modeling and diagnosing domain knowledge using Latent Semantic Indexing. *Proceedings of the 43rd Annual Meeting of the Human Factors and Ergonomics Society,* Houston, TX.

Gillan, D.J., & Cooke, N. J. (in press). Using Pathfinder networks to analyze procedural knowledge in interactions with advanced technology. In E. Salas (Ed.), *Human/Technology Interaction in Complex Systems.* Greenwich, CT: JAI Press Inc., Vol. 10.

Harman, D. (1986). An experimental study of the factors important in document ranking. In *Association for Computing Machinery Conference on Research and Development in Information Retrieval,* Association for Computing Machinery.

Holmes, M. (1997). Optimal matching analysis of negotiation phase sequences in simulated and authentic hostage negotiations. *Communication Reports, 10,* 1-8.

Howell, W. C., & Cooke, N. J. (1989). Training the human information processor: A look at cognitive models. In I. Goldstein (Ed.), *Training and Development in Work Organizations: Frontier Series of Industrial and Organizational Psychology, Volume 3* (pp. 121-182). New York: Jossey Bass.

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: MIT Press.

Jentsch, F. G., Sellin-Wolters, S., Bowers, C.A., & Salas, E. (1995). Crew coordination behaviors as predictors of problem detection and decision making times. In *Proceedings of the Human Factors and*

Ergonomics Society 39th Annual Meeting (pp. 1350-1353). Santa Monica, CA: Human Factors and Ergonomics Society.

Kiekel, P. A., Cooke, N. J., Foltz, P. W., Gorman, J., & Martin, M. (2002). Some promising results of communication-based automatic measures of team cognition. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting.*

Kleinman, D. L., & Serfaty, D. (1989). Team performance assessment in distributed decision making. *Proceedings of the Symposium on Interactive Networked Simulation for Training* (pp. 22-27). Orlando, FL: University of Central Florida.

Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management, 20,* 403-437.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 11-140

Landauer, T. K, Foltz, P. W. & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259-284.

Merritt, A. C., Helmreich, R. L. (1996). Human factors on the flight deck: The influence of national culture. *Journal of Cross-Cultural Psychology, 27,* 5-24.

Mosier, K. L., & Chidester, T. R. (1991). Situation assessment and situation awareness in a team setting. In Y. Quéinnec & F. Daniellou (Eds.), *Designing for everyone: Proceedings of the 11th Congress of the International Ergonomics Association* (pp. 798-800). London: Taylor & Francis.

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman and S. W. Draper (Eds.), *User centered system design* (pp. 31-61). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Orasanu, J. (1990). *Shared mental models and crew performance* (Report No. CSLTR-46). Princeton, NJ: Princeton University.

Oser, R. L., Prince, C., Morgan, B. B., Jr., & Simpson, S. (1991). An analysis of aircrew communication patterns and content (NTSC Tech. Rep. No. 90-009). Orlando, FL: Naval Training Systems Center.

Poole, M.S., Holmes, M., Watson, R., & DeSanctis, G. (1993). Group decision support systems and group communication: a comparison of decision making in computer-supported and nonsupported groups. *Communication Research, 20,* 176-213.

Rehder, B., Schreiner, M. E., Wolfe, B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes, 25*(2&3).

Salas, E., Bowers, C. A., & Cannon-Bowers, J. A. (1995). Military team research: 10 years of progress. *Military Psychology, 7,* 55-75.

Salas, E, Cannon-Bowers, J.A., Church-Payne, S., & Smith-Jentsch, K. A. (1998). Teams and teamwork in the military. In C. Cronin (Ed.), *Military Psychology: An Introduction* (pp. 71-87). Needham Heights, MA: Simon & Schuster.

Sanderson, P. M. & Fisher, C. (1994). Exploratory sequential data analysis: Foundations. *Human Computer Interaction, 9,* 251-317.

Schvaneveldt, R.W. (1990). Pathfinder associative networks: Studies in knowledge organization. Norwood NJ: Ablex.

Smith, J. B. (1994). *Collective Intelligence in Computer-Based Collaboration.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Stolcke, A., Reis, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics 26(3),* 339-373.

Stout, R. J. (1995). Planning effects on communication strategies: A shared mental model perspective. *Proceedings of the Human Factors Society 39th Annual Meeting* (pp. 1278-1282).

Sundstrom, E., DeMeuse, K. P., & Futrell, D. (1990). Work teams: Applications and effectiveness. *American Psychologist, 45,* 120-133.

Thornton, R. C. (1992). *The effects of automation and task difficulty on crew coordination, workload, and performance.* Unpublished doctoral dissertation, Old Dominion University, Norfolk, VA.

Watt, J. H. & VanLear, A. C. (1996). *Dynamic Patterns in Communication Processes.* Thousand Oaks, CA: Sage Publications.